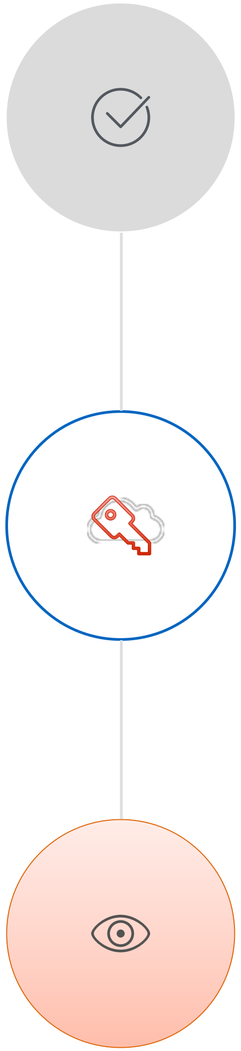
SPRINT7

目的はなにか



スクラッチを通してk-meansを理解する

クラスタ分析を行う

このスライドは？

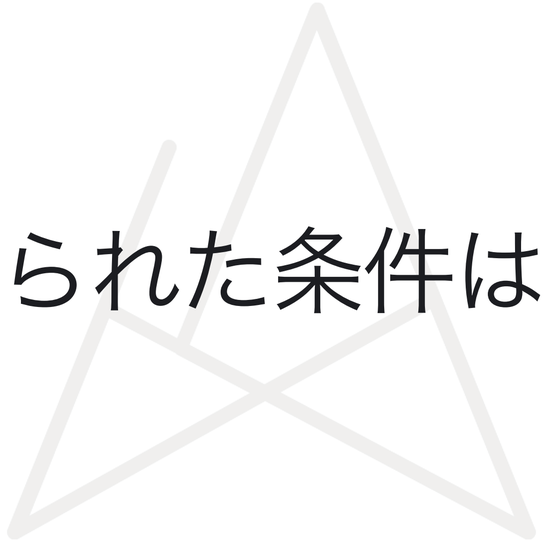
ここでは、k-Means法の

基本的な知識を学びましょう

k-Means法とはなにか

K-Meansアルゴリズムは、*k*個（固定数）の重心を識別し、重心位置を最適化する（平均値を最小化する）ために反復的な計算を行い、すべてのデータ点を最も近いクラスタ(1)に割り当てる仕組み。
クラスタ内の誤差平方和を削減することにより、すべてのデータポイントが各クラスタに割り当てられる。

(1) クラスターは、特定の類似性のために集約されたデータポイントの集合。



与えられた条件は何か

k-Means法においては以下が仮定されている。

① 入力データは特徴量行列Xのみ（教師なし学習）
② ハイパーパラメータとして固定値 k を入力する

この課題の対象者

① scikit-learnのクラスタリングモデルを用いて、学習、推定するコードが書ける方

この後の流れ

k-meansの幾何学的説明

① データ分布からランダムサンプリングしたk個のデータ点をクラスタの重心とする（kはハイパーパラメータ）
② 各重心に対しすべてのデータ点とのユークリッド距離を計算する
③ 各重心との距離が最小となるデータ点郡を、その重心に帰属するクラスタとする
④ k個のクラスタ毎にデータの平均となる点を求め、新しい重心とする
⑤ ②へ戻る

この後の流れ

実装上の手順を確認する

① サンプル数のインデックスに対し、kクラス分のランダムな初期ラベルを割り当てる
② 各ラベル毎にデータ点をグルーピングし、クラスタを作成する
③ クラスタ毎にデータ点の平均値を求め、そのクラスタの重心とする
④ その重心から、すべてのサンプルのデータ点との距離を計算する
⑤ 各データ点から見て、距離が最小となる重心のクラスタにそのデータ点を割り当てる
⑥ ③〜⑤を繰り返す
⑦ 収束条件（値が変化しない・定義した反復回数に達した等）を満たしたら、終了

SSEについて

クラスタ内**誤差平方和**（Sum of Squared Errors）
クラスタリングの性能評価関数。

関数

(データ点の座標)ー(重心座標)

$$SSE = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|X_n - \mu_k\|^2$$

自分が属するクラスタならば 1
自分が属さないクラスタならば 0